Práctica 1: Web scraping

Ainara Romero Roldan y Fidel Romero Zegarra

11 de abril de 2022

1. Contexto:

Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

Respuesta:

Desde el siglo XIX, la actividad humana ha sido el principal motor del cambio climático debido principalmente a la quema de combustibles fósiles, como el carbón, el petróleo y el gas, lo que produce gases que atrapan el calor. Estos gases son conocidos como Gases de Efecto Invernadero (GEI) y entre ellos principalmente se encuentran el Carbono Dióxido (CO_2) y Metano (CH_4). Como consecuencia, las temperaturas están aumentando radicalmente llegando a niveles preocupantes donde los patrones climáticos y equilibrio de la naturaleza se están viendo afectados.

En este contexto se pretende realizar una búsqueda de información en diferentes páginas web que permitan obtener un dataset para analizar los niveles de emisión de CO_2 y NH_4 por países, así como el índice de riesgo climático.

2. Título:

Definir un título que sea descriptivo para el dataset.

Respuesta:

Al final de la práctica se obtienen tres diferentes dataset [1] y la definición de un título descriptivo para cada uno sería el siguiente:

- Dataset 1: Índice de cambio climático por país en el año 2019.
- Dataset 2: Emisiones de Metano por país en el año 2018.
- Dataset 3: Emisiones de CO_2 por país desde 2001 al 2020.

[1] Se añade a cada dataset el link de su correspondiente página web, se determinarán también en el siguiente apartado

Se pretende en el apartado "Limpieza y análisis de los datos"juntar los diferentes dataset para obtener un único dataset que recoja toda la información. En ese caso, el título que definiría el dataset completo sería 'Emisiones de gases invernaderos por país' .

3. Descripción del dataset:

Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Respuesta:

Los dataset se han obtenido mediante el *web scraping* de las siguientes páginas web y contiene el siguiente contenido:

- Dataset 1: https://www.epdata.es/datos/cambio-climatico-datos-graficos/447

  El Índice de Riesgo Climático Global (IRC) indica el nivel de exposición y la vulnerabilidad a los fenómenos climáticos extremos. El dataset recoge el valor IRC del año 2019 de 161 países del mundo.

- Dataset 2: https://www.datosmundial.com/co2-por-pais.php

  El dataset muestra los 101 mayores productores de metano del mundo. Además, proporciona la información sobre emisiones de NH_4 por persona en cada país.

- Dataset 3: https://datosmacro.expansion.com/energia-y-medio-ambiente/emisiones-co2

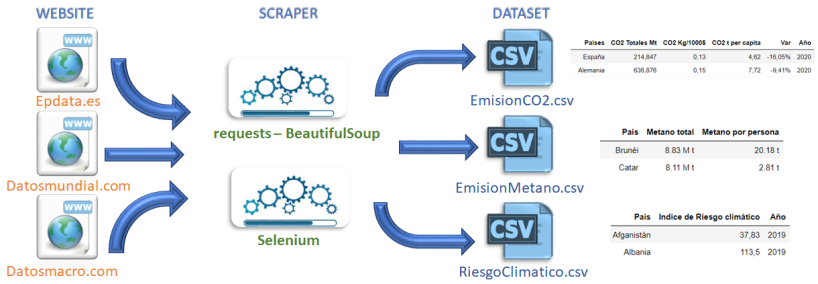
  El dataset almacena información referida a los niveles de emisión de CO_2 por cada país desde 2001 hasta el 2020. Además, se incluye información acerca de las emisiones totales, emisiones por cada 1000 dólares de PBI y emisiones per cápita.

4. Representación gráfica:

Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

Respuesta:

La siguiente representación muestra las fases del proyecto para obtener los diferentes datset:



5. Contenido:

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.

Respuesta:

Antes de iniciar el proceso de webscraping se han revisado los archivos *robots.txt* y el *sitemap* de cada página web con el objetivo de conocer las restricciones de cada sitio web y poder evitar las limitaciones de navegación para obtener la información requerida:

- Dataset 1:
  - *Robots.ts*
  - *Sitemap*
- Dataset 2:
  - *Robots.ts*
  - *Sitemap*
- Dataset 3:
  - *Robots.ts*
  - *Sitemap*

Los campos de cada dataset son los siguientes:

- Dataset 1:
  - **País**: país en el que se calculó el Índice de riesgo climático.
  - **Índice de riesgo climático**: valor de Índice de riesgo climático.
  - **Año**: año en el que se calculó el Índice de riesgo climático.

Los datos del dataset se han recogido el año 2019. Para el web scraping de esta página web se ha utilizado el lenguaje de programación Python en combinación con Selenium WebDriver. Esta herramienta permite simular la interacción de un usuario con cualquier navegador, en este proyecto se ha utilizado un servicio para Chrome y se ha hecho uso de esta herramienta debido a que la página a escrapear es dinámica y el dataset se obtuvo a partir de un gráfico dinámico.

- Dataset 2:

  - **País**: país en el que se calculó la emisión de CH_4.
  - **Metano total**: emisiones totales de metano. La cifra se da en el formato *cantidad Mt* donde *Mt* indica millones de toneladas.
  - **Metano por persona**: emisiones totales de metano por persona. La cifra se da en el formato *cantidad t* donde *t* indica toneladas.

  Los datos hacen referencia al año 2018. Para el *web scraping* de esta página web se ha utilizado el lenguaje de programación Python mediante las librerías Python Requests y BeautifulSoup. La primera nos permite realizar la descarga del sitio web que queremos escrapear mientras que la segunda librería nos permite obtener la estructura anidad de la página web y extraer la información de interés.

- Dataset 3:

  - **Año**: año en el que se calculó la emisión de CO_2.
  - **País**: país en el que se calculó la emisión de CO_2.
  - **CO_2 Totales Mt**: Emisiones totales de CO_2 de uso de combustibles fósiles y procesos industriales. *Mt* indica millones de toneladas.
  - **CO_2 kg/1000**: Emisiones de CO_2, por cada 1000 dólares del Producto bruto interno (PBI).
  - **CO_2 t per cápita**: Emisiones de CO_2 tonelada per cápita.
  - **Variación %**: Porcentaje de variación con respecto al año anterior

  Los datos hacen referencia desde el 2001 al 2020. Para el web scraping de esta página web se ha utilizado el lenguaje de programación Python en combinación con Selenium WebDriver. Esta herramienta permite simular la interacción de un usuario con cualquier navegador y se ha utilizado esta herramienta debido a que la página a escrapear es dinámica y al momento de navegar por los enlaces a los diferentes años inicialmente carga o una ventana emergente para cookies o una ventana emergente para dar información o publicitaria y Selenium facilita el cerrar estas ventanas para tener libertad de acceder a la información de interés.

6. Agradecimientos:

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Respuesta:

- Dataset 1: La información de este dataset fue obtenida de la página web de Epdata . El titular de este sitio web y por tanto propietario de de los datos es Europa Press Noticias S.A., y brinda acceso libre a parte de los datos que publican siempre que su uso no sea comercial de acuerdo con lo establecido en su "Aviso Legal": `https://www.europapress.es/avisolegal.html`
- Dataset 2: La información de este dataset fue obtenida de la página web de DatosMundial. El titular de este sitio web y por tanto propietario de los datos es Eglitis_media, y brinda acceso libre a los datos que publican de acuerdo a los términos establecidos en su "Protección de Datos": `https://www.datosmundial.com/privacidad.php`
- datDataset 3: La información fue obtenida de la página web de Expansión a través de su portal Datosmacro.com. El titular de este sitio web y por tanto propietario de los datos es Aldatanow, S.L, y brinda acceso libre a todos los datos que publican de acuerdo con lo establecido en sus "Términos de uso": `https://datosmacro.expansion.com/legal/terminos`

Es preciso indicar que se han realizado diversos análisis acerca del cambio climático, emisiones de gases invernadero y riesgo climático como se puede evidenciar en la página web de Statista:
`https://es.statista.com/temas/8615/el-cambio-climatico-a-nivel-mundial/#topicHeader__wrapper`
Asimismo, existen análisis publicados acerca de la emisión de CO_2 como los publicados por:

- Joseph Awonusi
- Sanskar Hasija

7. Inspiración:

Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Respuesta:

Nuestro principal objetivo es obtener datos que nos permitan realizar un análisis acerca de los niveles de emisión de CO_2 y NH_4 en el mundo, para conocer cuáles son los países más contaminantes. Asimismo, gracias a la data recolectada se podría realizar análisis aplicando diversos algoritmos de Machine Learning que permitan elaborar pronósticos o predicciones acerca de los niveles de contaminación por emisión de CO_2 a los que cada país puede llegar; También, se podría contrastar el índice de riesgo climático y las emisiones de CO_2 en el año 2019, incluso se podría comparar los niveles de emisiones de CO_2 y NH_4 y concluir cuál de ellos está generando mayor impacto. Los resultados podrían ayudar a tomar conciencia acerca del daño que se causa al planeta por la emisión de gases invernaderos y así los gobiernos propongan políticas de sostenibilidad.

8. Licencia:

Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Other (specified above).
- Unknown License.

Respuesta:

El tipo de licencia es **Released Under CC0: Public Domain License**. Decidimos optar por este tipo de licencia debido a que el tipo de información que se está generando es de mucha utilidad para cualquier persona y podría ayudar a que otros trabajadores de datos puedan realizar mayor procesamiento y así se podría generar más conocimiento acerca de la contaminación por emisiones de gases invernaderos.

9. Recursos

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Mitchel, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media, Inc. Chapter 1. Your First Web Scraper.

Contribuciones	Firma
Investigación previa	A.R.R., F.R.Z.
Redacción de las respuestas	A.R.R., F.R.Z.
Desarrollo del código	A.R.R., F.R.Z.